



US 20040117403A1

(19) **United States**(12) **Patent Application Publication**
Horn et al.(10) **Pub. No.: US 2004/0117403 A1**(43) **Pub. Date: Jun. 17, 2004**(54) **METHOD AND APPARATUS FOR QUANTUM CLUSTERING****Related U.S. Application Data**

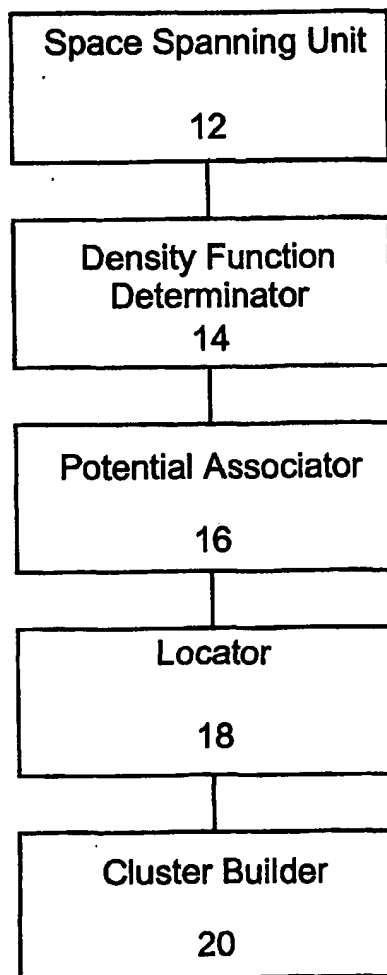
(60) Provisional application No. 60/290,385, filed on May 14, 2001.

(76) Inventors: **David Horn**, Tel Aviv (IL); **Assaf Gottlieb**, Hod Hasharon (IL); **Inon Axel**, Ramat Gan (IL)**Publication Classification**(51) **Int. Cl.⁷ G06F 17/00**(52) **U.S. Cl. 707/104.1**

Correspondence Address:

Anthony Castorina**G E Ehrlich****Suite 207****2001 Jefferson Davis Highway****Arlington, VA 22202 (US)**(57) **ABSTRACT**

A method of determining clusters of data within a dataset, the dataset is represented by a plurality of multidimensional data entries, the method comprises (a) spanning a space, represented by a plurality of points; (b) determining a density function over the space; (c) associating a potential to the density function; (d) locating a plurality of local minima of the potential; and (e) for each of the plurality of local minima, attributing at least one of the points; thereby determining clusters of data within the dataset.

(21) Appl. No.: **10/474,508**(22) PCT Filed: **May 14, 2002**(86) PCT No.: **PCT/IL02/00374**



US 20020183966A1

(19) **United States**(12) **Patent Application Publication****Mishra et al.**(10) **Pub. No.: US 2002/0183966 A1**(43) **Pub. Date: Dec. 5, 2002**

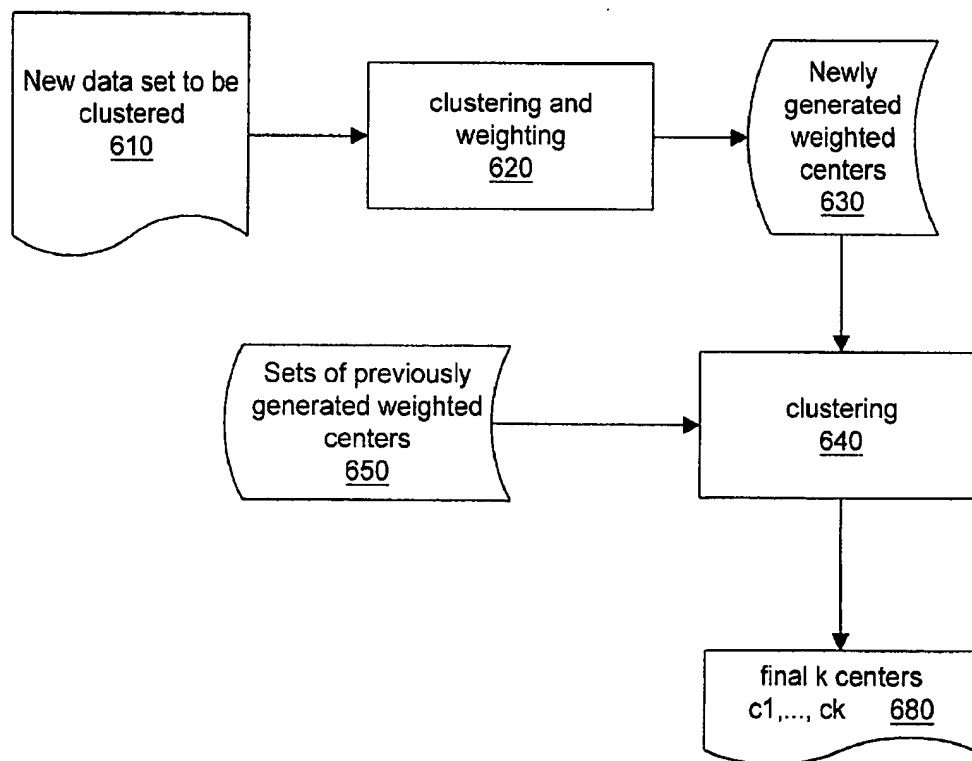
(54) **COMPUTER IMPLEMENTED SCALABLE,
INCREMENTAL AND PARALLEL
CLUSTERING BASED ON WEIGHTED
DIVIDE AND CONQUER**

Publication Classification(51) **Int. Cl.⁷ G06F 15/00**(52) **U.S. Cl. 702/179**

(76) **Inventors:** Nina Mishra, San Ramon, CA (US);
Lladan O'Callaghan, Mountain View,
CA (US); Sudipto Guha, Chatham, NJ
(US); Rajeev Motwani, Palo Alto, CA
(US)

Correspondence Address:**HEWLETT-PACKARD COMPANY****Intellectual Property Administration****P.O. Box 272400****Fort Collins, CO 80527-2400 (US)**(21) **Appl. No.: 09/854,212**(22) **Filed: May 10, 2001**(57) **ABSTRACT**

A technique that uses a weighted divide and conquer approach for clustering a set S of n data points to find k final centers. The technique comprises 1) partitioning the set S into P disjoint pieces S_1, \dots, S_p ; 2) for each piece S_i , determining a set D_i of k intermediate centers; 3) assigning each data point in each piece S_i to the nearest one of the k intermediate centers; 4) weighting each of the k intermediate centers in each set D_i by the number of points in the corresponding piece S_i assigned to that center; and 5) clustering the weighted intermediate centers together to find said k final centers, the clustering performed using a specific error metric and a clustering method A .





US006742003B2

(12) **United States Patent**
Heckerman et al.

(10) **Patent No.:** US 6,742,003 B2
(45) **Date of Patent:** May 25, 2004

(54) **APPARATUS AND ACCOMPANYING METHODS FOR VISUALIZING CLUSTERS OF DATA AND HIERARCHICAL CLUSTER CLASSIFICATIONS**

FOREIGN PATENT DOCUMENTS

WO	WO 90/04231	4/1990
WO	WO 95/31788	11/1995
WO	WO 95/34884	12/1995
WO	WO 96/28787	9/1996

(75) **Inventors:** David E. Heckerman, Bellevue, WA (US); Paul S. Bradley, Seattle, WA (US); David M. Chickering, Bellevue, WA (US); Christopher A. Meek, Kirkland, WA (US)

OTHER PUBLICATIONS

Brunk et al, "MineSet: An Integrated System for Data Mining", Data Mining and Visualization, (©1997, AAAI Press), pps. 135-138.

(73) **Assignee:** Microsoft Corporation, Redmond, WA (US)

(*) **Notice:** Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 548 days.

(List continued on next page.)

(21) **Appl. No.:** 09/845,151

Primary Examiner—Shahid Alam

(22) **Filed:** Apr. 30, 2001

(74) *Attorney, Agent, or Firm*—Amin & Turocy, LLP

(65) **Prior Publication Data**

US 2003/0018652 A1 Jan. 23, 2003

(57) ABSTRACT

(51) **Int. Cl.⁷** G06F 17/30

(52) **U.S. Cl.** 707/104.1; 707/4; 707/5;
707/10; 707/103 R; 704/202; 704/206;
705/26; 345/347

(58) **Field of Search** 707/3, 4, 5, 10,
707/103 R, 104.1; 705/5, 10, 23, 26; 704/9,
103, 522; 706/2; 701/200, 202, 206, 208;
345/347; 711/157, 173; 713/200; 709/201,
218, 219

(56) **References Cited**

U.S. PATENT DOCUMENTS

4,903,305 A	2/1990	Gillick et al.	381/41
5,506,986 A	4/1996	Healy	707/204
5,537,586 A	7/1996	Amram et al.	395/600
5,742,816 A	4/1998	Barr et al.	395/615
5,758,072 A	5/1998	Filepp et al.	395/200.5

(List continued on next page.)

A system that incorporates an interactive graphical user interface for visualizing clusters (categories) and segments (summarized clusters) of data. Specifically, the system automatically categorizes incoming case data into clusters, summarizes those clusters into segments, determines similarity measures for the segments, scores the selected segments through the similarity measures, and then forms and visually depicts hierarchical organizations of those selected clusters. The system also automatically and dynamically reduces, as necessary, a depth of the hierarchical organization, through elimination of unnecessary hierarchical levels and inter-nodal links, based on similarity measures of segments or segment groups. Attribute/value data that tends to meaningfully characterize each segment is also scored, rank ordered based on normalized scores, and then graphically displayed. The system permits a user to browse through the hierarchy, and, to readily comprehend segment inter-relationships, selectively expand and contract the displayed hierarchy, as desired, as well as to compare two selected segments or segment groups together and graphically display the results of that comparison. An alternative discriminant-based cluster scoring technique is also presented.

68 Claims, 29 Drawing Sheets

